4 Generalized cardinality-based bound

In (Koltchinskii, 2011) generalization bounds rely on L_2 diameter of δ-minimal sets. We use similar considerations to improve cardinality-based bound (3).

Lemma 4 *Suppose $A = A_1 \sqcup \cdots \sqcup A_T$ is a decomposition of classifier set A into disjoint sets. Then the following bound holds for pessimistic ERM:*

$$Q_\varepsilon(A) \leqslant \sum_{t=0}^{T} Q_\varepsilon(A_t). \tag{5}$$

Proof: Let $\mu(A, X)$ be a classifier, retrieved from set A by pessimistic ERM given training sample X. Using the definition of pessimistic ERM we get that $\delta(\mu(A_1 \sqcup A_2, X), X)$ is equal to either $\delta(\mu(A_1, X), X)$ or $\delta(\mu(A_2, X), X)$. This immediately implies that $Q_\varepsilon(A_1 \sqcup A_2) \leqslant Q_\varepsilon(A_1) + Q_\varepsilon(A_2)$, and the general case (5) follows by induction. ■

Lemma 5 *Let B be a set where all classifiers have equal number of errors. Consider arbitrary subset $A_t \subset B$. Then the following bound holds for pessimistic ERM :*

$$Q_\varepsilon(A_t) \leqslant Q_\varepsilon(B). \tag{6}$$

Proof: Let us notice that pessimistic ERM and discrepancy maximization are the same learning algorithms for sets where all classifiers have equal number of errors. For discrepancy maximization inequality (6) follows from the fact that function $f(A) = \max\limits_{a \in A} \delta(a, X)$ is monotonic in the sense that $A_1 \subset A_2$ implies $f(A_1) \leqslant f(A_2)$. ■

Previous lemmas allow one to split original set into clusters of classifiers with equal number of errors and then expand each of them to larger set with known overfitting bound. Below we give one example of such set:

$$B_r^m(a_0) = \{a \in A \colon \rho(a, a_0) \leqslant r, \text{ and } n(a, X) = m\}.$$

This set can be interpreted as the most compact set of classifiers with Hamming diameter $2r$ where all classifiers have equal number of errors.

Lemma 6 *Let $B \equiv B_r^m(a_0)$ be a set of classifiers defined above. Then*

$$Q_\varepsilon(B) = H_L^{\ell,\, m}\left(\tfrac{\ell}{L}(m - \varepsilon k) + \lfloor r/2 \rfloor\right) \cdot [m \geqslant \varepsilon k]. \tag{7}$$

Proof: Denote $s = n(a_0, X)$, $r' = \lfloor r/2 \rfloor$. Then

$$\min_{a \in B} n(a, X) = \begin{cases} 0, \text{ where } s \leqslant r', \\ s - r', \text{ where } s > r'. \end{cases}$$

Condition $\delta(a, X) \geqslant \varepsilon$ is therefore equivalent to $s \leqslant s_a(\varepsilon) + r'$, which implies (7). ■

Theorem 7 *Suppose $A = A_1 \sqcup \cdots \sqcup A_T$ is a decomposition of classifier set A into disjoint sets, such that within set A_t all classifiers make m_t errors each. Let $d_t \equiv \sup\limits_{a, a' \in A_t} \rho(a, a')$ denotes Hamming diameter of A_t. Then*

$$Q_\varepsilon(A) \leqslant \sum_{t=1}^{T} [m_t \geqslant \varepsilon k] \cdot H_L^{\ell, m_t}(s(\varepsilon) + \lfloor d_t/2 \rfloor). \tag{8}$$

Proof: follows immediately from lemmas 4, 5 and 6. ■

This bound can be much sharper than (3) because it accounts for similarities between classifiers.